Lucerne University of
Applied Sciences and Arts
HOCHSCHULE
LUZERN
Technik & Architektur
FH Zentralschweiz

Diplomand **Matthias Höfflin**
Dozenten Prof. Dr. Thomas Hunziker
Themengebiet Nachrichtentechnik & Signal Processing, Mechatronik, Automation & Robotik
Projektpartner Hochschule Luzern, Institut für Elektrotechnik

Modellfreie Regelung mit Reinforcement Learning

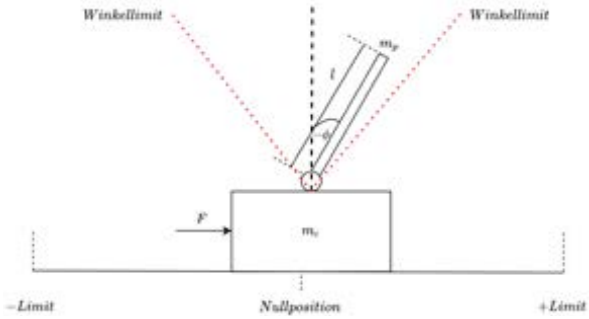

Abb. 1 Aufbau des Cartpole Versuchs

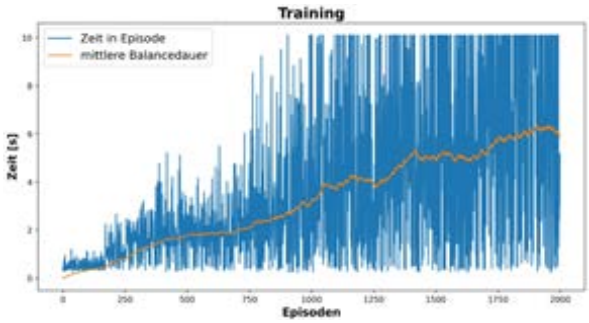

Abb. 2 Verhalten während des Trainings. Episodenabbruch bei Episodendauer > 10s oder Überschreitung der Limitierungen. Sichtbare Vergrösserung der mittleren Balancedauer.

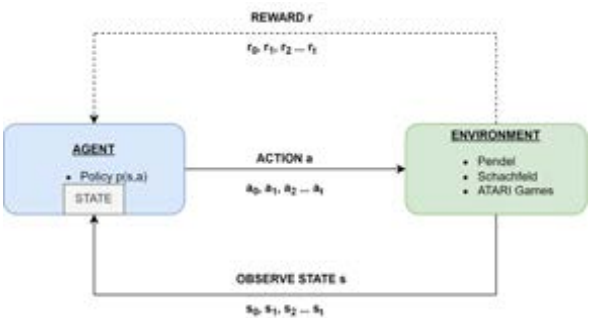

Abb. 3 Grundstruktur von Reinforcement Learning

Problemstellung

Einige Systeme aus der Regelungstechnik besitzen lediglich eine Zielfunktion, wie z.B. in Form einer Distanz vom gewünschten Zustand und keine weiteren Informationen über das Modell, welches zugrunde liegt. Reinforcement Learning soll über einen Agenten eine Policy erlernen und somit eine Umgebung meistern. Dabei kann eine explizite Modellierung der Systeme umgangen werden.

Lösungskonzept

Um eine geeignete Policy zu finden, wurde ein neuer Ansatz entwickelt. Der Ansatz wurde am Beispiel des räumlich inversen Pendels (Cartpole) untersucht und verfolgte die Grundidee, dass es zu jedem Zustand, der das Pendel einnehmen kann, eine Aktion besser ist als die andere. Ist dies der Fall, muss es möglich sein eine Entscheidungsgrenze in Form einer Hyperebene in die Zustandsdaten zu fitten, um zu jedem Zustand zu wissen, welche Aktion die geeignete ist. Die Untersuchung wurde an einem virtualisierten System des Versuchs ausgeführt.

Realisierung

Die Hyperebene wurde mithilfe von Support Vektor Maschinen (SVM) in die gesampelten Datenpunkte gefittet. Dabei wurde zusätzlich neben den vier Dimensionen des Pendels (Position des Wagens, Geschwindigkeit des Wagens, Winkel der Stange und die Winkelgeschwindigkeit) eine zusätzliche Reward-Dimension eingeführt. Der Reward beschreibt, wie gut eine Aktion war und ermöglicht somit eine Bewertung der Aktion. Diese Dimension ist notwendig, um eine Entscheidungsgrenze finden zu können. Die Entscheidungsgrenze wird in einem iterativen Sampling- und Trainingsverfahren nach und nach verbessert.

Ergebnisse

Der untersuchte Ansatz konnte durch iteratives Verbessern der Entscheidungsgrenze, mithilfe einer Support Vektor Maschine, das Pendel in den ausbalancierten Zustand bringen. Somit wurde die Grundidee bestätigt, dass es eine Tendenz in den Daten gibt, welche durch eine Hyperebene unterscheidbar ist.